MEMORY TECHNOLOGY AND APPLICATIONS

ALLEN RUSH
FELLOW, AMD
AGENDA

- Introduction
- Memory Resources in High Performance Compute Applications
- Challenges in Emerging Applications
- Future possibilities for memory-PE arrangements
Example: VR
- Frame rate, resolution, FOV, DR all contributing to higher capacity and BW
AUTOMOTIVE: INTELLIGENT VEHICLES ARE A REALITY

Throughput from sensors, recall from trained databases, big data updates...

Gaze detection, learning and prediction

Automotive safety: obstacles, cues, dictionary detection
PACKAGING AND MEMORY ARCHITECTURES TO MATCH HIGH DENSITY COMPUTE

HIGH BANDWIDTH MEMORY

- First in the Industry with High Bandwidth Memory (HBM) Technology
- 3D HBM DRAM Die Stack on Silicon Interposer
- >3X Performance/Watt Compared to GDDR5
- >50% Power Savings Versus GDDR5
HBM INTRODUCTION – SIZE, BW, POWER, INTERCONNECT: IMPROVEMENTS FOR APPLICATIONS ACROSS THE BOARD

AMD Radeon™ R9 Nano graphics card. Small size. Giant impact.

• The world’s first small form factor (6-inch) graphics card with High-Bandwidth Memory (HBM) delivering new advances in power efficiency.
• Powerful performance for unbelievably “real” 4K and VR gaming.
• A new paradigm for the Mini-ITX PC.
REAL TIME VISION AND AUDIO PRESENCE
PHOTOREALISTIC AND FULL 3D ACOUSTICS FOR GAMING, VR, AR

To achieve the vision of full presence you need

Scalable CPUs, GPUs, Accelerators, High BW Memory

Sensory Integration

Other Senses
ROOFLINE PERFORMANCE MODEL
MATCHING BW TO PEAK COMPUTE PERFORMANCE

- Peak performance depends on application memory density and BW
- Arithmetic intensity – ratio of actual ops (add, mul...) to mem ops (loads, stores)
- Applications with high AI: maximize peak performance

Roofline Design – Matrix kernels
- Dense matrix multiply
- Sparse matrix multiply

Graph showing performance models and operational intensity.
DNN rapidly becoming reliable solution for Machine Learning

- CNN and derivatives are contemporary forms that produce accurate results for many classification problems

- Memory BW and capacity
  - Different for Forward (Inference) and Backward (training)
  - For Training: 1M+ image training data base; 100-1000 parameters to train
  - For Inference: 4x 30fps 4K video ->~1B pixels/s; parameters fixed, variable data
ML PROGRESSION

Early ANN
Simple NN models

Backward Propagation
Supervised learning

Handwriting recognition
Deep Belief Networks

CNN and Deep NN
Accuracy
Image Classification
"Train to Learn"

Complex (very) deep
NN, combinations of RNN, RCNN,
LSTM...
Real time
Inference
Fast Training

Present Day

Real-time scene interp
and residual learning
"Learn to Train"
DEEP LEARNING APPLICATIONS
AND MANY MORE

- Immersive Gaming: Learned Behavior
- Medicine: Learning to interpret scans
- Log in Authentication: Learning pose and shape
- Vision Classification and Recall
- Security
- Advanced Driver Assistance: “Thinking Cars”
FUTURE REQUIREMENTS
CREATIVE WAYS TO CONSTRUCT DENSE COMPUTE-MEMORY MODULES

- Hard to tell: BW, capacity, power advances – limit?
- Need to get PEs close to L3 or PM
- Novel approaches: LIM

Logic-in-Memory
BIG DATA REQUIREMENTS

- Increasing size and demand for data access-speed of retrieval and storage
- DL NN structures with ~1B parameters and >100M input frames
- Sort, match and classify in near real time
- New training or inference data uploaded at ~2000GB/s
Factors driving advanced memory designs
- Applications requiring real time video, VR, advanced graphics

Increased CPU/GPU performance
- Need for balancing BW, capacity
- HBM solutions

Emerging applications with unique memory requirements
- ML – training and inference

Novel solutions for PE-Mem structures

Big Data
- More data upload
- Cloud DL: massive parameter and training data sets